Prediction of 2024 US election ...*

Colin Sihan Yang Lexun Yu Siddharth Gowda

November 3, 2024

We forecast the winner of the 2024 US presidential election using "poll-of-polls" by building a linear model.

1 Introduction

Election result forecasting has become an essential tool for analysts in political science and the public to predict the outcome of democratic process, such as the presidential election in the United States. Traditionally, individual polls have been used as a snapshot of voter sentiment, but they only reflect temporary changes in the performance of contestants, instead of a precise estimation of the election result. As discussed by Pasek (2015) and Blumenthal (2014), the aggregation of multiple polls, or "poll-of-polls," has become a popular technique to reduce individual survey errors and provide more accurate election forecasts. However, the traditional poll aggregation does not reflect dynamics of an election, especially with real-time changes and the introduction of new data. This creates a gap for a more adaptable model to predict the election result based on both polling data and additional variables, such as historical data and economic indicators.

This paper fills the gap by building a hybrid election forecasting model following the strategies mentioned by Pasek (2015). As Pasek (2015) described in their article, aggregation involves determining which surveys are worth including, as well as selecting, combining and averaging results from multiple polls to reduce individual biases and errors. Prediction modeling adds other data to the model that predicts election outcomes based on current dynamics. Hybrid models like the Bayesian approach incorporates prior beliefs based on historical data or expert knowledge and new evidence like economic updates to dynamically adjust the forecast as the campaign progresses.

In this paper, we aim to predict the 2024 us election result with the hybrid election forcasting model. We incorporate aggregation by filtering the polls on FiveThirtyEight (2024) by

*Code and data are available at: https://github.com/yulexun/uselection.

numeric grade that indicates pollster's reliability, prediction that incorporates social and economic indicators including unemployment rates and abortion rates, and hybrid approaches that leverages Bayesian techniques which combines historical data such as the 2016 election data, allowing for a dynamic prediction of the U.S. presidential election.

The estimand for this research paper is the predicted support percentages for Kamala Harris and Donald Trump. The prediction is based on quantifying various polling factors, including sample size, poll scores, and transparency scores, which are used as predictors.

The results of this model indicate a more stable and accurate forecast compared to traditional aggregation methods alone, [update this …]

The remainder of this paper is structured as follows: [update this …]

2 Data

2.1 Overview

For the data we used in this analysis about the polling result for Kamala Harris and Donalad Trump in 2024 USA president election.
- **response variable:** pct(pct: The percentage of the vote or support that the candidate received in the poll)
- **numeric predictor:**
sample size(sample_size: The total number of respondents participating in the poll)
timegap(the time gap between the poll start date and the real election date i.e timegap = real US election date - poll start date)
pollscore(A numeric value representing the score or reliability of the pollster in question)
- **categorical predictor** state(The U.S. state where the poll was conducted or focused)
methodology(The method used to conduct the poll)

2.2 Measurement

In this dataset, each row represents a polling question that records the variables of interest. Each entry allows us to explore the real-world relationships between polling factors and the support percentage (`pct`) for the candidates Kamala Harris and Donald Trump. This dataset enables an analysis of how various polling characteristics influence the reported support levels for the candidates we are focused.

2.3 Clean Data

The data cleaning process involves several steps to ensure the quality and relevance of the polling data. First, we filter the dataset to retain only poll results with a numeric grade of 2.7 or higher, indicating that the polls are considered reliable. Next, we address missing values in the state attribute: polls with NA in the state column are considered national polls.

We then create a new attribute, days_taken_from_election, which represents the time gap between the poll's start date and the actual U.S. election date. Additionally, we filter the dataset to include only polls conducted after July 21, 2024, the date when Kamala Harris declared her candidacy. Finally, we remove any remaining rows that contain missing values to ensure a clean dataset.

Table 1: Sample of cleaned US election data

pct	sample_size	pollscore	days_taken_from_election	state	methodology	Candidate Name
47.6	4180	-0.8	24	National	Online Ad	Kamala Harris
50.7	4180	-0.8	24	National	Online Ad	Donald Trump
0.8	4180	-0.8	24	National	Online Ad	Jill Stein
0.1	4180	-0.8	24	National	Online Ad	Chase Oliver
0.1	4180	-0.8	24	National	Online Ad	Cornel West
48.1	4180	-0.8	24	National	Online Ad	Kamala Harris

2.3.1 States included in anaylsis

After the data cleaning process, 21 states had no polling data. A table showing the number of polls for each state, including those without any polls, is provided in Table 2.

This absence of polling data is not a significant concern due to the structure of the United States Electoral College (explained in detail in the Appendix). The states lacking polling data have consistently followed historical voting patterns, so predicting the winning candidate in those states is unnecessary

2.4 Basic Statistics Summary for Data

In figure Figure 1, historically Democratic are polling for Kamala and historically Republican states are polling for Trump. Similarly, historically swing states also appear to be close, for

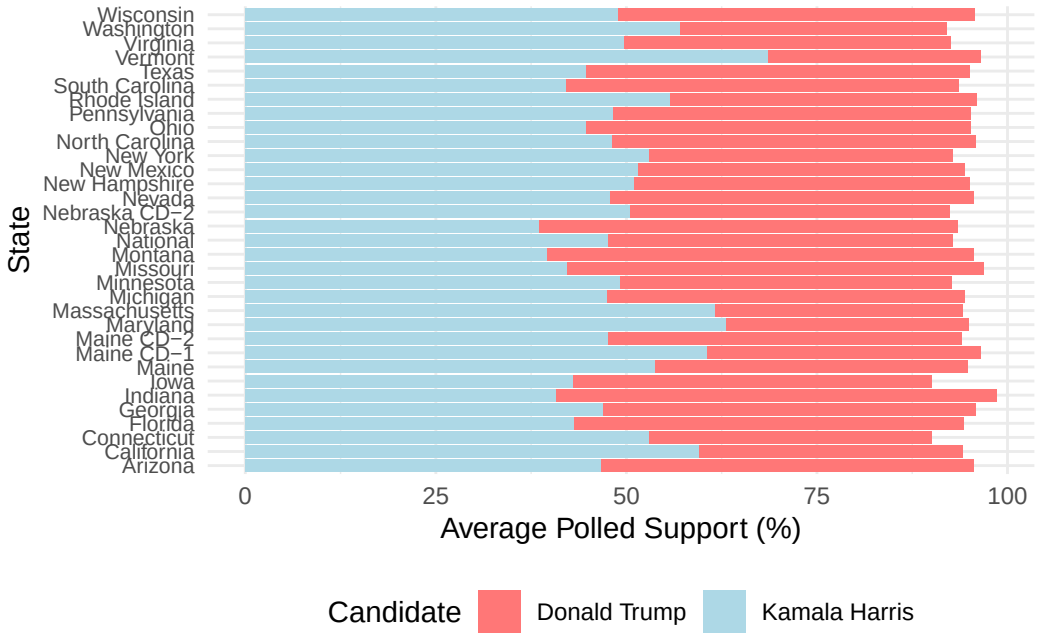

Figure 1: Historical State Voting Trends Are Maintined in 2024

instance Michigan (46.9% Trump, 47.5% Harris), Nevada (47.7% Trump, 47.9% Harris), and Pennsylvania (46.9% Trump, 48.2% Harris) all are close to an even split.

Figure Figure 2 shows that polls utilizing multiple communication methods to reach voters tend to have lower interquartile ranges (IQRs) in their boxplots compared to those that rely on only one or two communication methods.

2.5 Relationship Between Variables

Figure Figure 3 illustrates a weak positive correlation between a pollster's pollscore and the sample size of their poll. The figure also shows that most polls have a sample size around 800 to 1200 participants. It is also important to note that a few polls with exceptionally large sample sizes were excluded from the graph due to their status as clear outliers.

Based on Figure 4, there does not seem to be a relationship between the sample size of a poll the percentage of a candidate.

Figure **?@fig-pollscore-candidate-pct** depicts the relationship between a pollster's poll score and a candidate's percentage. For Donald Trump, a negative correlation is observed, indicating that pollsters with lower poll scores tend to assign him a higher percentage of support. Conversely, Kamala Harris shows the opposite trend: as the poll score decreases, her percentage tends to rise. This suggests that more reliable polls, characterized by lower poll scores, report higher support for Trump compared to less reliable pollsters.

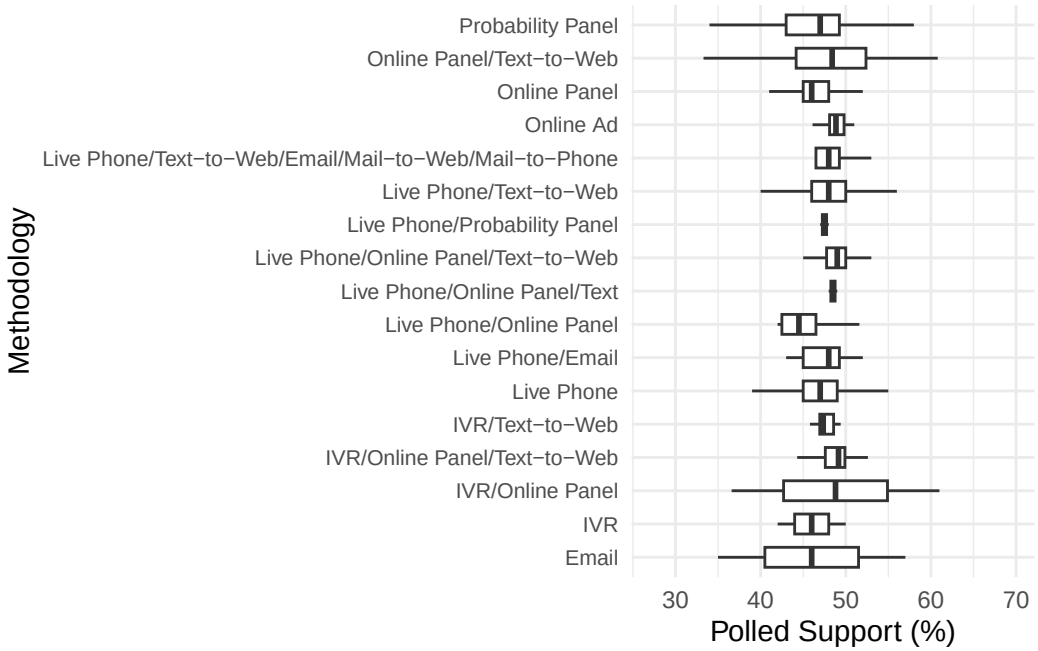

Figure 2: Polls with more Methodologies have Less Variability.

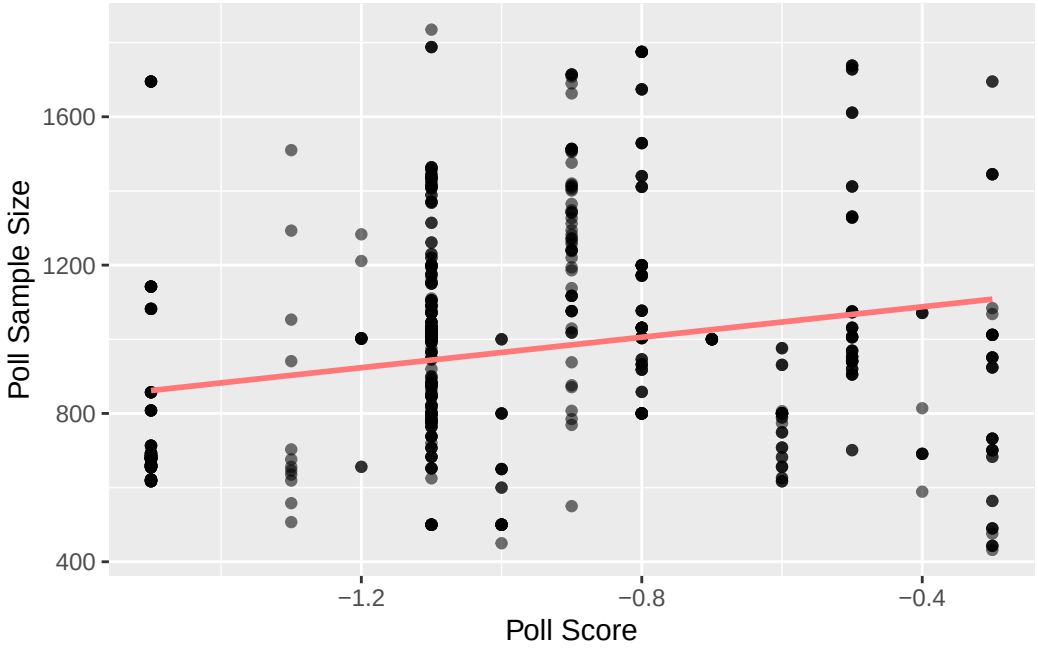

Figure 3: More Reliable Pollsters Have Larger Sample Sizes in their Polls

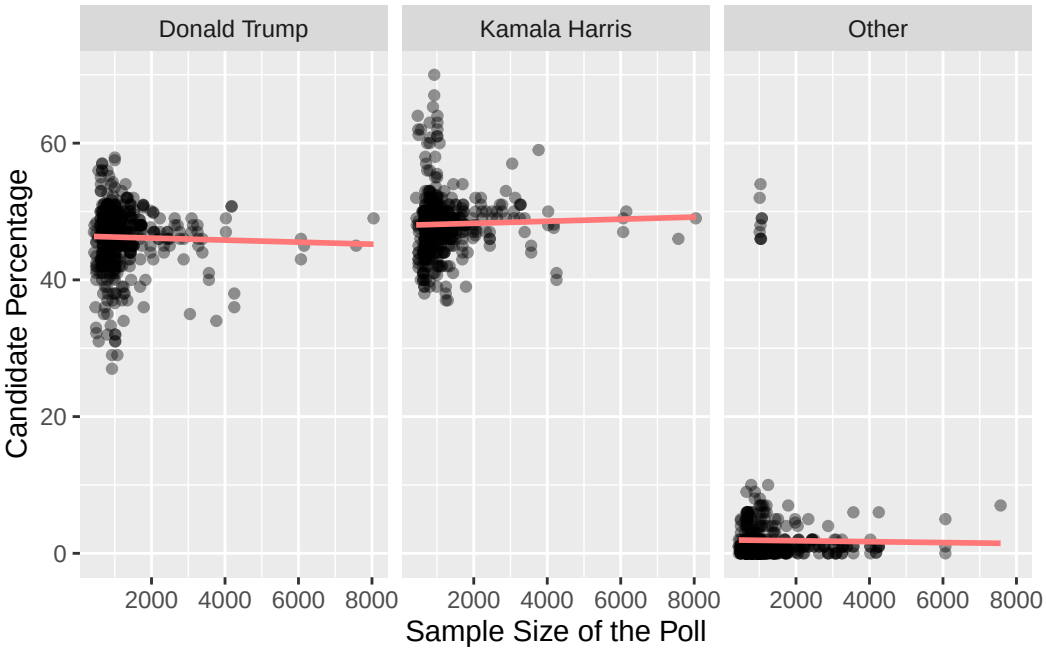

Figure 4: The Sample Size of a Poll does not impact a Candidate's Voting Percentage

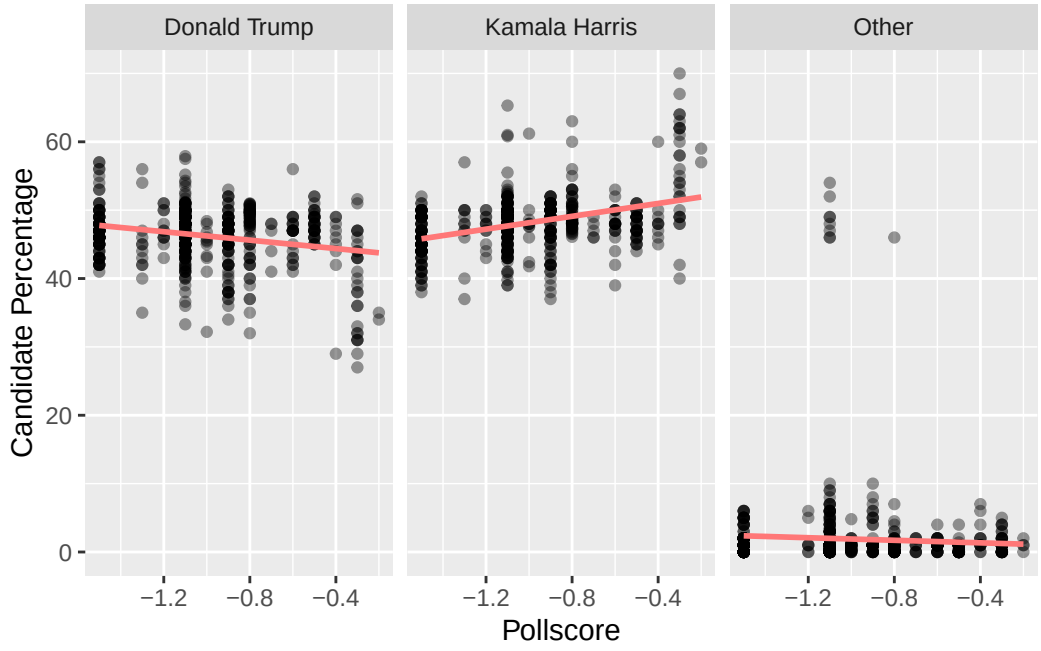

Figure 5: More Reliable Pollsters Score Higher For Trump

Based on **?@tbl-state-candidate-support**, state might have to be a removed variable.

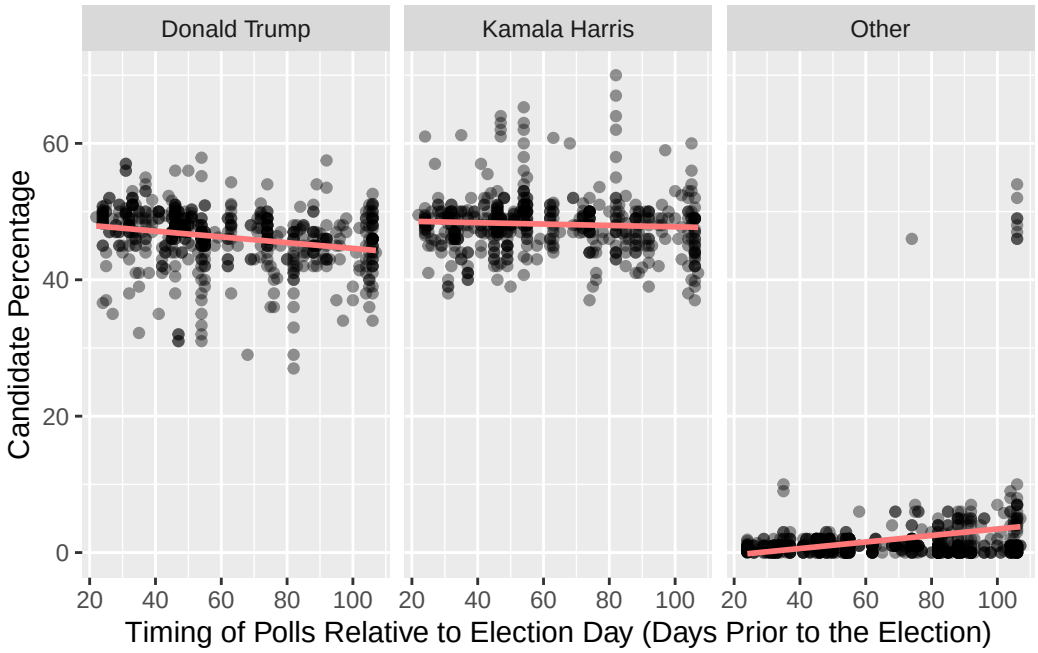

Figure 6: Both Canidate Percentages Are Gaining Support Closer to the Election

Based on Figure 6, Trump and Harris are getting more votes in polls that are done closer to the election. This is a result of non-major candidate support rapidly decreasing. Specifically, Trump support is increasing at a faster rate than Harris. However, the polls in general show a slight lead for Harris throughout the last 100 days.

The rapid decline in third party support could be due to Robert F. Kennedy dropping out of the race (this sentences should probably be in results or discussion section).

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`. us

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

C FiveThirtyEight Licenses

C.1 Overview of American Electon System

C.2 Brief Description of American Federal Goverment

The American federal or national government is split into three branches, executive., legislative, and judical The executive branch includes the president and the military. The legislative branch include two subgroups, the House of Representatives and the Senate. These two subgroups create laws. Every state has two Senators and one Representative per approximately 750,000 people. The judicial branch is court system.

C.2.1 What is the Electoral College

The Electoral College is the system used in the United States to elect the president and vice president. Instead of a direct popular vote, each state is allocated a certain number of electors based on its representation in Congress (the total of its Senators and Representatives). When voters cast their ballots, they are actually voting for a slate of electors pledged to a candidate. The candidate who receives a majority of electoral votes (270 out of 538) wins the presidency. This system means that winning the popular vote in a state generally results in winning all of that state's electoral votes. The only exception are the states of Maine and Nebraska, who award electoral votes by congressional district, with two additional votes given to the statewide winner.

The Electoral College results in some states being unnecessary to campaign in, as their strong historical voting pasterns towards either Democrats or Republicans make them unlikely to change, regardless of campaign efforts. Therefore, for statisticians, polling information from these states may not be that useful when trying to predict the outcome of an election. On the other hand, states that can vote either Democratic or Republican (swing states) are immensely important when predicting an election. As a result, campaigns spend hundreds of millions of dollars campaigning and understanding voters there.

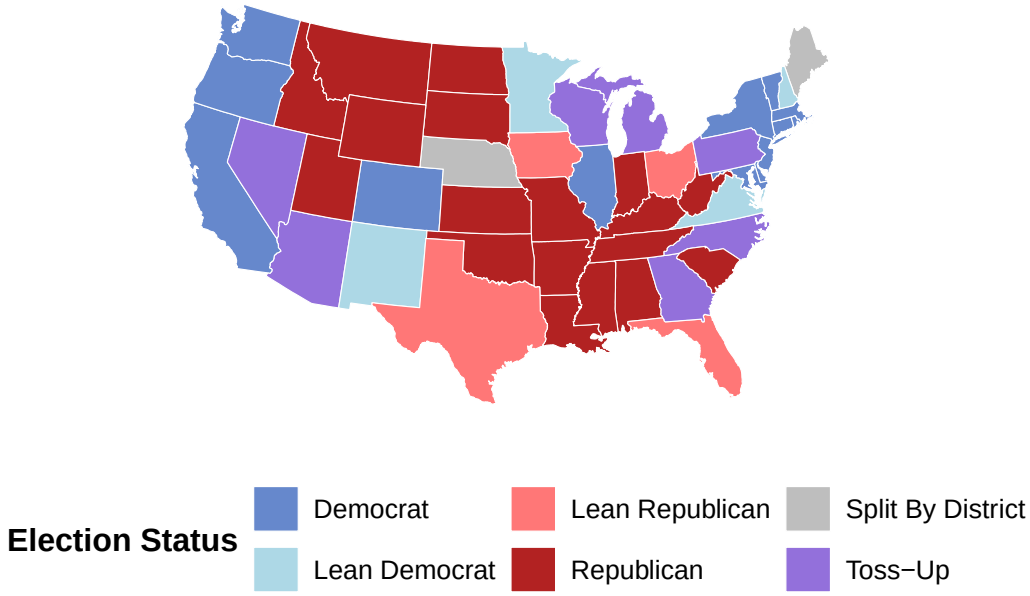

Figure 7: 2024 U.S. Presidential Election State Forecast Map

The state statuses presented in this map are based on evaluations from (**cnn?**), (**foxnews?**), and (**msnbc?**), which are generally agreed upon within the American political community. These organizations assessed historical voting patterns and recent polling data to derive their conclusions.

Notably, Nebraska and Maine are indicated in gray due to their delegates being split by district. In Nebraska, the state overall is projected to lean Republican, with the first and third districts also strongly favoring Republican candidates, while the second district leans Democratic. In Maine, the overall expectation is a Democratic leaning, consistent with its first district, though the second district leans Republican. Furthermore, Alaska and Hawaii are not in the map. Alaska is strongly favoring Republicans while Hawaii strongly favors democrats.

The seven generally agreed upon swing states are Pennsylvania, North Carolina, Georgia, Arizona, Nevada, Wisconsin, and Michigan with Texas, Florida, Nebraska District Two, Maine District 2, and Minnesota as the next closest races.

C.3 States Poll Count

Table 2: Polls included in Analysis Per State

State	Number of Polls In Analysis
Alabama	0
Alaska	0
Arizona	17
Arkansas	0
California	5
Colorado	0
Connecticut	1
Delaware	0
Florida	6
Georgia	17
Hawaii	0
Idaho	0
Illinois	0
Indiana	1
Iowa	1
Kansas	0
Kentucky	0
Louisiana	0
Maine	2
Maryland	2
Massachusetts	3
Michigan	18
Minnesota	5
Mississippi	0
Missouri	2

Table 2: Polls included in Analysis Per State

State	Number of Polls In Analysis
Montana	3
Nebraska	2
Nevada	9
New Hampshire	5
New Jersey	0
New Mexico	2
New York	3
North Carolina	23
North Dakota	0
Ohio	4
Oklahoma	0
Oregon	0
Pennsylvania	28
Rhode Island	2
South Carolina	1
South Dakota	0
Tennessee	0
Texas	6
Utah	0
Vermont	1
Virginia	4
Washington	1
West Virginia	0
Wisconsin	20
Wyoming	0
National	77

C.4 The Ideal Survey

C.4.1 Objective

The research team has developed a survey and distribution methodology with a hypothetical budget of $100,000. This objective is to create a polling system that accurately predicts the 2024 United States Election.

C.4.2 Sampling Frame

Given the monetary constraint, the team's first decision is to mostly include swing states and districts. These are Nevada, Arizona, Wisconsin, Michigan, Pennsylvania, North Carolina, and Georgia. Swing states are the primary focus of the poll because they disproportionately affect the outcome of the election based on the explanation in **?@sec-PLACEHOLDER**. However, the poll will also include Texas, Florida, and Minnesota, Nebraska District 2, and Maine District 2, as they have the potential to one by either party but are not as likely to change as the previously mentioned swing states. Nevertheless, it is important to note that to save costs, the team will focus on polling true swing states compared to Texas, Florida and Minnesota.

C.4.3 Survey Sending Process

The team will then find and use multiple databases, such as (**uspostalserviceresidentialaddressfile?**), to find addresses, telephones, and emails of potential voters in those states. According to (**pewresearch?**), gathering this information reduces non-response and selection bias, as the team can contact the same individual through multiple mediums. Moreover, certain demographics may prefer specific communication forms; for instance, the elderly may prefer phone or paper mail polls over text or email. Additionally, to encourage individuals to complete the survey, one in every 50 participants will have a chance at winning $20.

C.4.4 Sampling Methodology

Each state will have its own poll, but the polling methodology and the survey given will remain the same. Specifically, the polls will use stratified random sampling to make sure that participants reflect each state's counties in proportion to their population sizes. Ideally, the stratification would also ensure race, gender, age, and other socioeconomic factors are also accounted for. However, these factors are almost impossible to determine while sending the survey. Therefore, the research team decided to ask demographic questions directly inside the survey. After data collection, the team will weigh data based on demographics. For example, if a certain county has a 30% black population (this statistic will be determined from the US census), but only 15% of the survey participants are black, then the pollsters may decide to count each black participant's responses twice.

Furthermore, the team will aim to sample approximately 1,000 people from each true swing state. This sample size is large enough to provide reliable conclusions but not so large that it resembles sampling with replacement. For true swing states, the team aims to survey 1,000 individuals. However, if the final sample size falls short of this target, it is not considered a significant issue.

C.4.5 Survey Implementation & Question Creation

The team's ideal survey will be made using (**qualtrics?**). It will attempt to ensure four things: no leading questions, no question order bias, no answer order bias, and the survey should identify non-engaged participants. To identify participants who are not engaged, the research team has decided to add a worthless question. This question is extremely simple and clearly has one correct answer. Therefore, if a participant selects the wrong answer, they most likely are not engaged with the survey and their responses should be removed from the final data. An example of this is question number 7.

Order bias occurs when the sequence of questions subconsciously influences participants' responses. To prevent this, many questions should be randomized upon entry to the survey. However, some questions must follow a specific order, such as question 11, while others like questions 1-4 can be randomized. The team's hypothetical survey has not implemented this feature, though it is available with the (**qualtrics?**) paid plan.

Answer bias, like order bias, occurs when the order of answer choices influences participants' selections, with the first option often chosen more frequently. To prevent this, answer choices should be randomized upon survey entry. Questions 8 & 9 could benefit from this. Likewise, the team's hypothetical survey has not implemented this feature, but it is available in the (**qualtrics?**) paid plan.

A leading question occurs when a question is written in a way that suggests the user to give a certain answer. For example, "given that children are the future of our country, should we invest more money in their education". To prevent this, the researchers have ensured questions are written in a style where no unnecessary details or opinions are added.

C.4.6 Survey Questions

Click this (link)[https://qualtricsxm7d2hxss4j.qualtrics.com/jfe/form/SV_1Tu3PT2eUEa1Op8] for the (**qualtrics?**) survey.

List of all of the questions:

1. Select your race(s) (racial options where chosen based on (**whitehouseracialoptions-PLACEHOLDER?**))
    - White
    - Black or African American
    - American Indian or Alaska Native
    - Native Hawaiian or Pacific Islander
    - Middle Eastern or North African
    - Asian
    - Other
    - Prefer not to answer

2. Please select your gender

    - Male
    - Female
    - Non-binary
    - Prefer not to say

3. Please enter your age (in numbers)

- This is a text input field. Please note that this field has the auto-validation feature set to numbers in (**qualtrics?**). As a result, participants can only input numbers in this field and are alert if they have not.

4. Please select the highest degree of education you have obtained

    - GED Certificate
    - High School Diploma
    - Undergraduate Degree
    - Graduate Degree (Masters/Phd)
    - None
    - Other
    - Prefer not to answer

5. Are you a registered voter for the 2024 United States Presidential Election?

    - Yes
    - No

6. Place yourself on the political spectrum

    - Far Left
    - Center Left
    - Center
    - Center Right
    - Far Right

7. Can pigs fly?

    - Yes
    - Maybe
    - No
    - Prefer Not To Say

8. What political party have you registered with?

    - Republicans
    - Democrats
    - Green Party

- Libertarian
- Other
- Independent (unregistered)

9. Who will you vote for in the 2024 presidential election?

    - Democrat - Kamala Harris
    - Republican - Donald Trump
    - Green Party - Gill Stein
    - Libertarian - Chase Oliver
    - Other
    - Will not vote

C.4.7 Potential Problems with the Methodology And Polls

While the team's methodology and survey creates a robust system, there are potential issues. The reliance of weighting results based on a candidate's demographics can lead to error propagation. For instance, if a certain racial demographic population is only captured limitedly and that limited sample is far from representative, a few individuals in the population can have a large impact on the polls prediction of what candidate will win the state. There is also a selection bias in terms of the monetary reward. Potentially, people who like monetary rewards could be more likely to engage in the survey and could therefore exhibit certain voting or demographic characteristics that create an unrepresentative sample.

References

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science and Politics* 47 (2): 297–300. http://www.jstor.org/stable/43284537.

FiveThirtyEight. 2024. "Our Data." *FiveThirtyEight*. https://data.fivethirtyeight.com.

Pasek, Josh. 2015. "THE POLLS–REVIEW: PREDICTING ELECTIONS: CONSIDERING TOOLS TO POOL THE POLLS." *The Public Opinion Quarterly* 79 (2): 594–619. http://www.jstor.org/stable/24546379.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.